

Discrete Choice Modeling with PROC MDC

John Ang, AutoZone, Memphis TN

ABSTRACT

Discrete choice modeling involves the prediction of an individual's response given several distinct alternatives, where one and only one decision must be made. While the ability to analyze such models has long been available in SAS®, the development of PROC MDC enables the SAS user to directly address the issue. This paper will demonstrate how to generate the input data set for PROC MDC, utilize PROC MDC to implement discrete choice modeling, and interpret the results.

INTRODUCTION

Prior to the introduction of PROC MDC, discrete choice modeling was largely implemented in SAS® using PROC PHREG (Proportional Hazards Regression) or SAS procedures involving binomial or multinomial logistic regression (e.g. PROC LOGISTIC). The multinomial logit model was fit using PROC PHREG with the TIES=BRESLOW option, as the likelihood function of the resulting survival analysis model has the same form as that of the multinomial logit model. However, the use of PROC PHREG required some preliminary data processing to convert the data into the appropriate format.

PROC MDC (Multinomial Discrete Choice), a relatively new SAS procedure in the SAS/ETS® software, analyzes models where the choice set consists of multiple alternatives. This procedure supports conditional logit, mixed logit, heteroscedastic extreme value, nested logit, and multinomial probit models. It uses the maximum likelihood (ML) or simulated maximum likelihood method for model estimation. While some data pre-processing may be needed to create the required input data set, PROC MDC is specifically geared towards the modeling of discrete choice situations, and is the most intuitive choice (in this author's estimation) among SAS procedures for this type of analysis.

MODEL FORMULATION

The multinomial logit model is used to model relationships between a polytomous response variable and a set of regressor variables. Applied to discrete choice modeling, the underlying analytical model for PROC MDC is described as follows:

Let m be the number of alternatives under consideration. Also, let each alternative be described by a set of n independent explanatory variables. Then, we can postulate the following utility function for each alternative:

$$U_i = e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni}} \quad (1)$$

where U_i = utility of i th alternative, $i = 1, \dots, m$
 x_{kj} = value of k th explanatory variable for j th alternative, $j = 1, \dots, m$, $k = 1, \dots, n$
 β_k = linear parameter for k th explanatory variable, $k = 1, \dots, n$

The probability P_i of choosing the i th alternative is calculated as follows:

$$P_i = \frac{U_i}{\sum_{j=1}^m U_j} \quad (2)$$

It must be noted that the utility function in (1) assumes constancy in preferences among individuals. In general, utility values for the same alternative will vary among different individuals. Nevertheless, this mathematical model assumes that one can estimate values of β_k , $k = 1, \dots, n$, that would describe the utility preferences of the "average" decision-maker.

THE MDC PROCEDURE

PROC MDC can be utilized to estimate the values of β_k , $k = 1, \dots, n$, for the n regressor variables. However, the input SAS dataset for PROC MDC requires that there be as many observations as there are alternatives, for each decision maker. The ID statement is required for all MDC models—it identifies the variable that refers to the decision maker, which must also be present in the input SAS dataset for PROC MDC. Finally, there must be a variable that indicates the alternative chosen by each decision maker. The value must be 1 for the actual alternative chosen, and one and only one observation per decision maker can have this value.

Creating The Input SAS Dataset

Consider a SAS dataset with the following layout:

```
store  qty1  qty2  qty3  x11  x12  x13  . . . . . x $n$ 1  x $n$ 2  x $n$ 3
```

This SAS dataset contains unit sales data for 3 substitutable items, all SKUs of a particular product but with different quality levels. For each store in the retailer's network, the variables qty1, qty2 and qty3 denote the amount of unit sales for the product of good quality, the product of better quality, and the product of best quality, respectively. The x_{ij} variables contain the values of the explanatory variables, where x_{ij} is the value of the n th explanatory variable for the product of j th quality ($j=1$ refers to good, $j=2$ better and $j=3$ best).

Furthermore, let us assume that in buying one unit of a particular item, the decision-maker was, in effect, opting not to buy the other two items of different quality levels. Referring to the dataset shown above as the SAS work dataset storesales, we can use the following SAS code to create the input SAS dataset for PROC MDC:

```
%macro setxvars(ptype);
x1 = x1&ptype;
x2 = x2&ptype;
.....
x $n$  = x $n$ &ptype;
%mend setxvars;

data trans(keep=store subj option sold x1 x2 ..... x $n$ );
  set storesales;
  retain subj 0;
  do cnt=1 to qty1;
    subj + 1;
    option = 1; sold = 1; %setxvars(1); output trans;
    option = 2; sold = 0; %setxvars(2); output trans;
    option = 3; sold = 0; %setxvars(3); output trans;
  end;
  do cnt=1 to qty2;
    subj + 1;
    option = 1; sold = 0; %setxvars(1); output trans;
    option = 2; sold = 1; %setxvars(2); output trans;
    option = 3; sold = 0; %setxvars(3); output trans;
  end;
  do cnt=1 to qty3;
    subj + 1;
    option = 1; sold = 0; %setxvars(1); output trans;
    option = 2; sold = 0; %setxvars(2); output trans;
    option = 3; sold = 1; %setxvars(3); output trans;
  end;
run;
```

If desired, some variable transformations can also be performed.

```
data trans;
  set trans;
  /* VARIABLE TRANSFORMATIONS */
  SAS code;
run;
```

Running PROC MDC

The work dataset trans is now ready for PROC MDC!

```
proc mdc data=trans outest=est_mdc;
  model sold = x1 x2 ..... xn
           / type=clogit choice=(option 1 2 3);
  id subj;
  output out=out_mdc p=fcst xbeta=sumxb;
run;
```

The variable sold is the dependent variable, while x_1 , x_2 , ..., x_n are the explanatory variables. The TYPE=CLOGIT option indicates that the conditional logit model is being analyzed. The alternative is reflected in the variable option, of which there are three possible values: 1, 2, and 3. The decision-maker is indicated by the variable subj—all observations with the same value of subj refer to the same purchaser.

The OUTEST statement designates est_mdc as the SAS dataset that will hold the values of β_k , $k = 1, \dots, n$. The OUTPUT statement specifies out_mdc as the SAS dataset that will contain the output of this PROC MDC procedure. This output dataset will also contain a variable fcst whose value is the predicted probability that the decision-maker designated by subj will choose the alternative designated by option, as well as a variable sumxb that holds the linear predictor values, i.e. the inner product of β_k , $k = 1, \dots, n$, with the explanatory variables x_1 , x_2 , ..., x_n .

The SAS Online Documentation provides more information on the syntax for PROC MDC.

Interpreting The Results of PROC MDC

Once the values of β_k , $k = 1, \dots, n$, have been estimated, the discrete choice model described by (1) and (2) can be further analyzed to derive solutions and instructive insights. To do so, we need to expand on this model.

This utility function asserts that an alternative's utility value is only dependent on the values of its own set of explanatory variables, i.e.

$$\frac{\partial U_i}{\partial x_{ki}} = \beta_k U_i \quad \frac{\partial U_i}{\partial x_{kj}} = 0 \text{ for } i \neq j$$

For that matter, it is also assumed that the explanatory variables are independent not only within an alternative, but also among the other alternatives:

$$\frac{\partial x_{ki}}{\partial x_{kj}} = 0 \text{ for } i \neq j \quad \frac{\partial x_{li}}{\partial x_{kj}} = 0 \text{ for } l \neq k$$

The partial derivatives of P_i with respect to x_{ki} , $j = 1, \dots, m$, are calculated as follows:

$$\begin{aligned}
 \frac{\partial P_i}{\partial x_{ki}} &= \frac{1}{\left(\sum_{j=1}^m U_j\right)^2} \left[\left(\sum_{j=1}^m U_j\right) \frac{\partial U_i}{\partial x_{ki}} - U_i \frac{\partial \left(\sum_{j=1}^m U_j\right)}{\partial x_{ki}} \right] \\
 &= \frac{1}{\left(\sum_{j=1}^m U_j\right)^2} \left[\left(\sum_{j=1}^m U_j\right) \beta_k U_i - U_i \beta_k U_i \right] \text{ since } \frac{\partial \left(\sum_{j=1}^m U_j\right)}{\partial x_{ki}} = \frac{\partial U_i}{\partial x_{ki}} = \beta_k U_i \\
 &= \beta_k \left(\frac{U_i}{\sum_{j=1}^m U_j} \right) \left(\frac{\sum_{j \neq i} U_j}{\sum_{j=1}^m U_j} \right) \\
 &= \beta_k P_i (1 - P_i) \quad (3)
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial P_i}{\partial x_{kl}} &= \frac{1}{\left(\sum_{j=1}^m U_j\right)^2} \left[\left(\sum_{j=1}^m U_j\right) \frac{\partial U_i}{\partial x_{kl}} - U_i \frac{\partial \left(\sum_{j=1}^m U_j\right)}{\partial x_{kl}} \right] \\
 &= \frac{1}{\left(\sum_{j=1}^m U_j\right)^2} (-U_i \beta_k U_l) \text{ since } \frac{\partial U_i}{\partial x_{kl}} = 0 \text{ for } i \neq l \\
 &= -\beta_k P_i P_l, \quad i \neq l \quad (4)
 \end{aligned}$$

Note that for $i \neq l$, $\frac{\partial P_i}{\partial x_{kl}} = \frac{\partial P_l}{\partial x_{ki}} = -\beta_k P_i P_l$.

The elasticity ε_{ki} of choosing the i th alternative with respect to its k th explanatory variables is calculated as

$$\begin{aligned}
 \varepsilon_{ki} &= \frac{\partial P_i}{\partial x_{ki}} \cdot \frac{x_{ki}}{P_i} \\
 &= \beta_k P_i (1 - P_i) \left(\frac{x_{ki}}{P_i} \right) \\
 &= \beta_k x_{ki} (1 - P_i) \quad (5)
 \end{aligned}$$

Similarly, the elasticity ε_{kil} of choosing the i th alternative with respect to the k th explanatory variable for the l th alternative is calculated as

$$\begin{aligned}\varepsilon_{kil} &= \frac{\partial P_i}{\partial x_{kl}} \cdot \frac{x_{kl}}{P_i} \\ &= -\beta_k P_i P_l \left(\frac{x_{kl}}{P_i} \right) \\ &= -\beta_k x_{kl} P_l, \quad i \neq l\end{aligned}$$

Note that $\varepsilon_{kil} = -\beta_k x_{kl} P_l$ does not depend on i . Thus, we can refer to ε_{kil} as follows:

$$\varepsilon'_{kl} = -\beta_k x_{kl} P_l, \quad i \neq l \quad (6)$$

The expected value of the k th explanatory variables is calculated as follows:

$$E(x_k) = \sum_{j=1}^m P_j x_{kj}$$

Each of the addends in the above equation, i.e. $P_j x_{kj}$, can be thought of as the contribution of an alternative to $E(x_k)$. The partial derivatives of $P_j x_{kj}$ with respect to x_{kj} , $j = 1, \dots, m$, are calculated as follows:

$$\begin{aligned}\frac{\partial(P_i x_{ki})}{\partial x_{ki}} &= P_i \left(\frac{\partial x_{ki}}{\partial x_{ki}} \right) + x_{ki} \left(\frac{\partial P_i}{\partial x_{ki}} \right) \\ &= P_i + x_{ki} \beta_k P_i (1 - P_i) \\ &= P_i (1 + \beta_k x_{ki} (1 - P_i)) \\ &= P_i (1 + \varepsilon_{ki}) \quad (7)\end{aligned}$$

$$\begin{aligned}\frac{\partial(P_i x_{ki})}{\partial x_{kj}} &= P_i \left(\frac{\partial x_{ki}}{\partial x_{kj}} \right) + x_{ki} \left(\frac{\partial P_i}{\partial x_{kj}} \right) \\ &= x_{ki} (-\beta_k P_i P_j) \quad \text{since} \quad \frac{\partial x_{ki}}{\partial x_{kj}} = 0 \quad \text{for} \quad i \neq j \\ &= P_j \varepsilon'_{ki} \quad (8)\end{aligned}$$

CONCLUSION

PROC MDC can be utilized to estimate the values of the linear parameters in the multinomial logit model described by equations (1) and (2). Some preliminary data processing may be required to create the appropriate input dataset for PROC MDC.

In this paper, we examined the mathematical formulation of the underlying model and discussed the general format of the input SAS dataset for PROC MDC (along with ways to create that dataset from other SAS datasets). The last section presented some quantitative results that could be utilized to derive solutions and insights. If nothing else, the solution of the linear parameters using PROC MDC allows the SAS user to input this discrete choice model, along with appropriate constraints, into an optimization solver like SAS/OR®.

REFERENCES

Kuhfled, Warren F. and So, Ying (1996) "Multinomial Logit Models", presented at SUGI 20 and published in its proceedings

SAS Institute Inc. (2002-2006) SAS Online Doc[®] 9.1.3, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

John Ang
AutoZone
123 South Front Street, Dept 9013
Memphis TN 38135
(901) 495-6064 office
(901) 495-8434 fax
john.ang@autozone.com
www.autozone.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.